Data Mining at FDA

Hesha J. Duggirala (CVM), Joseph M. Tonning (CDER), Ella Smith (CFSAN), Roselie A. Bright (OITI), John D. Baker (CVM), Robert Ball (CBER), Carlos Bell (CDER), Khaled Bouri (OCS), Susan J. Bright-Ponte (CVM), Taxiarchis Botsis (CBER), Marc Boyer (CFSAN), Keith Burkhart (CDER), G. Steven Condrey (ORA), James J. Chen (NCTR), Stuart Chirtel (CFSAN), Ross W. Filice (OCS), Henry Francis (CDER), Hongying Jiang (CDRH), Jonathan Levine (OITI), David Martin (CBER), Taiye Oladipo (CFSAN), Rene O'Neill (CDRH), Lee Anne M. Palmer (CVM), Antonio Paredes (CTP), George Rochester (CTP), Deborah Sholtes (CTP), Hui-Lee Wong (CDRH), Zhiheng Xu (CDRH), Ana Szarfman (CDER), Taha Kass-Hout (OITI)

ABSTRACT

This article summarizes past and current data mining activities at FDA. We address data miners in all sectors, anyone interested in the safety of products regulated by FDA (predominantly medical products, food, veterinary products and nutrition, and tobacco products), and those interested in FDA activities. Topics include routine and developmental data mining activities, short descriptions of the mined FDA data, advantages and challenges of data mining at FDA, and future directions of data mining at FDA.

INTRODUCTION

The diverse products regulated by the U.S. Food and Drug Administration (FDA) represent approximately 25% of the U.S. economy, are used daily, and affect the health of many millions of people and animals. Besides food and drugs, they include, for example, nutritional supplements, genetically engineered foods, vaccines, artificial hearts, surgical lasers, devices used to administer drugs and biologics, gene therapies, veterinary drugs, pet food, tobacco products, and many others. Adverse events associated with these products are responsible for tremendous public health and financial costs. These adverse-event-related costs impact healthcare product development, health insurance premiums, and healthcare services (e.g., hospitalizations) all of which lead to long-term societal losses such as permanent disability and death. [1] Ensuring the safety of these products is a formidable challenge.

FDA collects and maintains data that provide safety information for its regulated products, the largest being databases of reports of safety problems presumed to be associated with marketed products. The annual number of reports received has steadily increased over the decades due to factors such as increases in population, the number and type of regulated products, awareness of the importance of reporting, and increased ease with which reports can be submitted (e.g. online tools). The FDA currently receives approximately two million adverse event, use error, and product complaint reports each year from consumers, health care professionals, manufacturers,

and others. These reports are entered into various databases maintained by the FDA for subsequent analyses to identify potential safety issues and enhance the understanding of those issues. Since the 1990s, FDA has been exploring and expanding its use of data mining to:

- Cope with the increasing sizes of the reports databases
- Speed identification of potential safety issues
- Aid in prioritizing potential safety issues, and
- Free personnel to devote a higher proportion of their time to tasks that aren't yet readily assisted by machines.

As basic data mining methods have become routine for more and more safety report databases, FDA has recommended its use to the drug industry [2] and FDA data mining experts have expanded their attention to adding more sophisticated data mining methods and applying data mining to other types of product safety-related FDA and non-FDA databases.

In this paper we summarize the current data mining tools and methods the FDA uses to identify safety signals. We also address the expansion of data mining to include new types of methods and to address additional databases. This FDA webpage incorporates input from all of the FDA regulatory Centers in addition to the FDA's Office of the Commissioner (OC), which serves all the Centers. More details about FDA's organization are available here.

DATA MINING METHODS APPLIED TO SAFETY REPORTS

FDA's larger databases of safety reports are analyzed with routine and prototype data mining methods and tools.

Disproportionality methods

As applied at FDA, disproportionality methods are largely used to identify statistical associations between products and events in their respective databases of safety reports. Such methods compare the observed count for a product-event combination with an "expected" count. Unexpectedly high reporting associations "signal" [3] that there **may** be a causal association between the particular adverse event and the product. Identified safety signals are referred to as Disproportionately Reported Combinations (DRCs).

The Proportional Reporting Ratio (PRR) is the foundational concept for many disproportionality methods. [4, 5] PRR is the degree of disproportionate reporting of an adverse event for a product of interest compared to this same event for all other products in the database. Thus, the entire database is used as a background "expected." The PRR relies on an "independence assumption," i.e., that there is no association between products and events mentioned in reports. If there is disproportionate reporting of an event for a particular product, then this independence

assumption is questionable, i.e., there may in fact be an association between the particular adverse event and the product. Such an "association" is statistical; thus, it cannot be interpreted as causal or related to risk.

This concept of disproportionality may be displayed by means of a contingency table where

- "a" is all reports for a *specific* adverse event ("Event Y") for Product (e.g. a drug) X,
- "b" is all reports for *all other* adverse events for Product X,
- "a + b" are *all* the reports for Product X,
- "c" is all reports for *all other products* for Event Y,
- "d" is all reports for *all other products* for *all other adverse event*, and
- "c + d" is all reports for all other products.

Table 1. PRR Contingency Table

	Event Y	All other events	
Product X	a	b	a + b
All other products	c	d	c + d
	a + c	b + d	Total

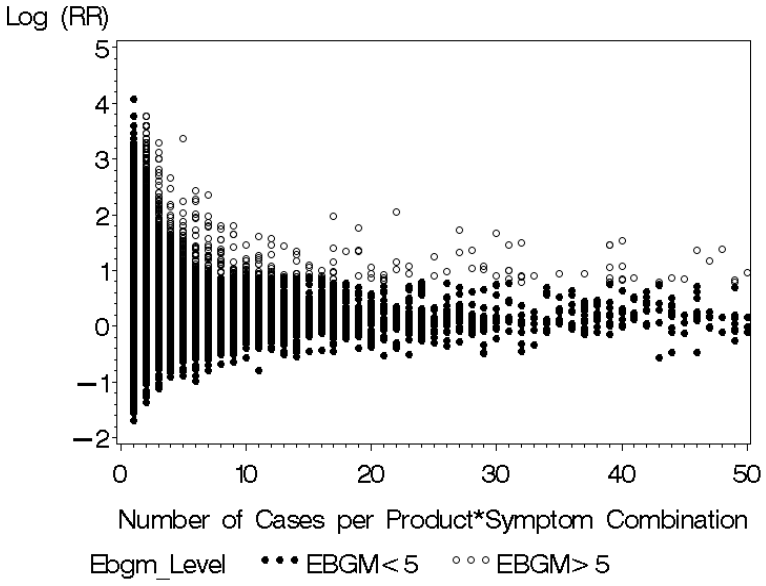
The PRR = [a/(a+b)] / [c/(c+d)]. Finney [4] and Evans [5] explored disproportionate adverse event reporting, and this concept is the basic foundation for various data mining methods the FDA currently uses. If the ratio of [a/(a+b)] is greater than the ratio of [c/(c+d)], then Event Y is "disproportionately reported" for Product X, when the rest of the database is considered as a background "expected." PRR is a method that does not adjust for multiplicity, small counts, or the underlying fact that every report represents a suspicion of an adverse event related to a product.

However, because this method does not adjust for small observed or expected numbers of reports of the product-event pair of interest, other more advanced statistical methods are employed, such as the Multi-Item Gamma Poisson Shrinker (MGPS), [2, 6, 7] which produces Empirical Bayesian Geometric Mean (EGBM) scores. The EBGM calculation is conceptually similar to that of the PRR, but incorporates Bayesian "shrinkage" and stratification to produce disproportionality scores toward the null, especially when there are limited data and small numbers of cases. One important difference between the PRR and EBGM estimates is that in the case of PRR the adverse events from the product in question do not contribute to the number of "expected" cases, while all adverse events from the product contribute to the expectation when using EBGM. The statistical modifications used in the EBGM methodology diminish the effect of spuriously high PRR values, thus reducing the number of false-positive safety signals. [6] Thus, EBGM values provide a more stable estimate of the relative reporting rate of an event for a particular product relative to all other events and products in the database being analyzed. [7] Lower and upper 90% confidence limits for the EBGM values are denoted EB05 and EB95, respectively.

Several FDA Centers including CDER, CBER, and CFSAN, use the MGPS algorithm for their routine surveillance activities. Various commercially available software programs generate PRR and/or EBGM scores (e.g., Empirica Signal™, PVAnalyser™, MASE™ and SAS™). NCTR has used EBGM scores applied to FAERS data to improve drug-induced liver toxicity prediction. In addition, NCTR has used a bi-clustering data mining algorithm with pattern recognition techniques for analysis of FAERS data.

The reason why product-event combinations with small numbers of reports must be "shrunk" is made apparent in the following plot of the log10 reporting ratio (RR) vs the number of reported cases for the product-event combination. The RR represents the ratio of the number of observed cases to expected cases under the assumption of independence between products and symptoms.[8]

Figure 1: Variance of log(RR)

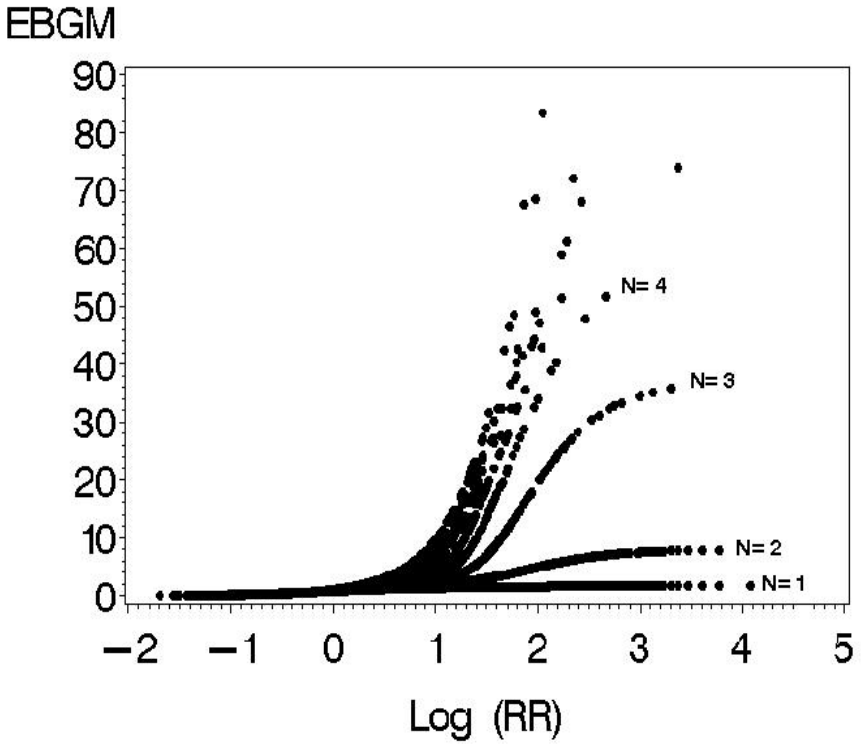


The variance of log(RR) decreases rapidly with increasing number of cases per product-event combination (Figure 1). The open circles in the graph represent product-event combinations with EBGM scores greater than 5, which are clearly reporting signals. Note that with only two cases the first open circle is evident where log(RR) equals 4 or RR equals 10,000. With 10 cases per product-event combination open circles appear at log(RR) equals 2 or absolute RR equals 100. With 20 cases, open circles appear at log(RR) equals 1 or RR equals 10. Thus, a much lower RR is required to generate a high EBGM score with a greater number of cases.

The extreme shrinkage of EBGM with low numbers of cases is demonstrated in the "squid-like" plot below (Figure 2), where at n=1 a log(RR) of 4 (RR=10,000) is reduced to an EBGM score approaching 1. Slightly less shrinkage occurs at n=2; by the time n= 4, it is much easier to generate clear signals in terms of EBGM scores. As n (number of cases per product-event cases)

increases in the plot from 5 to 485 cases per product-event combination, the data can be adequately represented by a single cubic curve. [9]

Figure 2: Extreme shrinkage of EBGM



Change-point analysis (CPA)

Change-point analysis (CPA) is a statistical method for determining whether a change in either the slope [10, 11, 12, 13, 14, 15] or variability [16] has taken place in a time series or sequence in very large databases. Detecting such changes is important in many different application areas such as economics, medicine, agriculture, and intelligence. Kass-Hout, et al. [10, 11] applied CPA to the active syndromic surveillance data to detect changes in the incidence of Emergency Department (ED) visits due to daily Influenza-Like-Illness (ILI) during the H1N1 pandemic. The development of CPA methods starts from the single change point detection before extending to multiple change points. The detection of a single change point can be posed as a hypothesis test. The null hypothesis, H_0, corresponds to no change point (m = 0) and the alternative hypothesis, H_a, is a single change point (m = 1). The general likelihood-ratio- based approach and Cumulative Sum (CUSUM) method were widely used to detect a change in the mean within normally distributed observations. As the volume of time series data increases, there is a growing need to maintain situation awareness and be able to efficiently and accurately estimate the

location of multiple change points. Several multiple changepoint algorithms have been developed such as Binary Segmentation, [12] Segment Neighbourhoods [13, 14] and the Pruned Exact Linear Time (PELT). [15] In addition to the changes in means problems, researchers have developed a new CPA method that detects changes in variance to search for changes in variability. [16] As a complimentary tool to the signal detection efforts at FDA, CPA could be critical for public health regulation, surveillance of adverse events and recalls, and regulators' understanding of the longitudinal effect of adverse events from their regulated products.

Text mining

Text mining is of interest due to the volume of data submitted in adverse event reports that is "unstructured" (e.g. narratives, event descriptions).

A recently developed text mining system, Vaccine *a*dverse *e*vent Text Mining (V*ae*TM) system, was derived in CBER from the vaccines adverse event reports database. V*ae*TM currently extracts diagnostic, treatment, and various assessment features using rules. [17] The newest version of the system (released in summer 2014) includes laboratory test results and temporal information modules; the latter associates the above features on a time axis and provides a critical overview of the adverse events following the administration of not only vaccines but also drugs.

Incorporation of reference data into data mining

For drugs, CDER is evaluating and advising product development for a proprietary software tool called Molecular Analysis of Side Effects (MASE™). [18] MASE™ integrates the publicly available adverse drug event reports data with various chemical and biological data sources in a drug-centric manner. This software tool is being utilized to assess the biological plausibility of safety signals. The program can identify targets, enzymes, and transporters that are disproportionately associated with drugs and events. This "mechanism mining" tool generates enzymatic, pathway, and molecular target hypotheses that warrant further evaluation. The program was recently used to study infusion reactions. [18]

Beyond the Office of Crisis Management's experiences with geographical information systems (GIS) technology to manage product quality threats due to natural disasters, [19] FDA is also exploring GIS technology to enable safety data analysis for routine circumstances. Product surveillance using GIS will allow analysts to capture, store, retrieve, analyze, manage, and display safety data geographically and/or temporally. Tracking potential safety signals in this manner can provide new opportunities for real-time interventions, and identification of:

- populations at risk (e.g., those with genetic predispositions to specific events),
- identification of patterns related to intentional or unintentional product contamination

- identification of areas where public health education and assistance may be appropriate.

Visualization tools

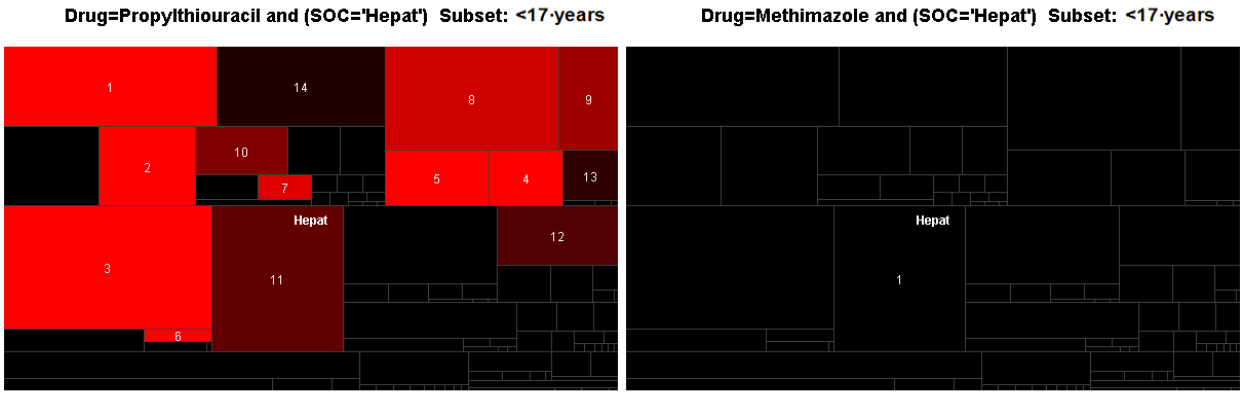
CBER uses the network analysis (NA) technique, which incorporates automated pattern recognition and has been applied to VAERS. [20] Another prototype tool, Adverse Event Network Analyzer (AENA), [21] incorporates various algorithms to identify patterns in VAERS data and is now equipped with other functionalities to support the processing of other types of data.

Regardless of the analytical tools used, visualization of the data is paramount. Because of the volume and complexity of the data, extremely helpful graphical tools used at FDA include heat maps [22] and sector maps. [23] Visualization tools that place related products and related outcomes near each other, [23] and that can also display contrasting sub-groups, [22] are very valuable. The following two examples show the process of examining the EBGM values associated with hepatotoxicity in children and adolescents treated with either propylthiouracil or methimazole. CDER scientists used interactive graphic tools tailored to improve understanding of the signals in the safety data.

Propylthiouracil and methimazole have been used interchangeably for the treatment of Graves' disease in pediatric patients for over 60 years. Major differences were observed in the number and proportion of severe liver injury reports for propylthiouracil compared with methimazole.

Figure 3 shows a comparative sector map of the MGPS data mining profile within the Hepatic System Organ Class in individuals age 17 or younger. This display groups preferred terms using a dictionary of medical terms, allowing the identification of redundant signals for a drug and the visual comparison of these signals across drugs. The hepatic safety profiles of propylthiouracil and methimazole are shown side by side. The sector map for propylthiouracil shows strong and redundant signals for several adverse event terms associated with severe hepatotoxicity. These safety effects were not seen with methimazole.

Figure 3: Comparative sector map (heat map) display of adverse events in the MedDRA Hepatic System Organ Class for propylthiouracil (left) and methimazole (right) in individuals less than 17 years of age. The numbers indicate the ranking of Empirical Bayesian Geometric Mean (EBGM) values for each preferred term (PT). [23]

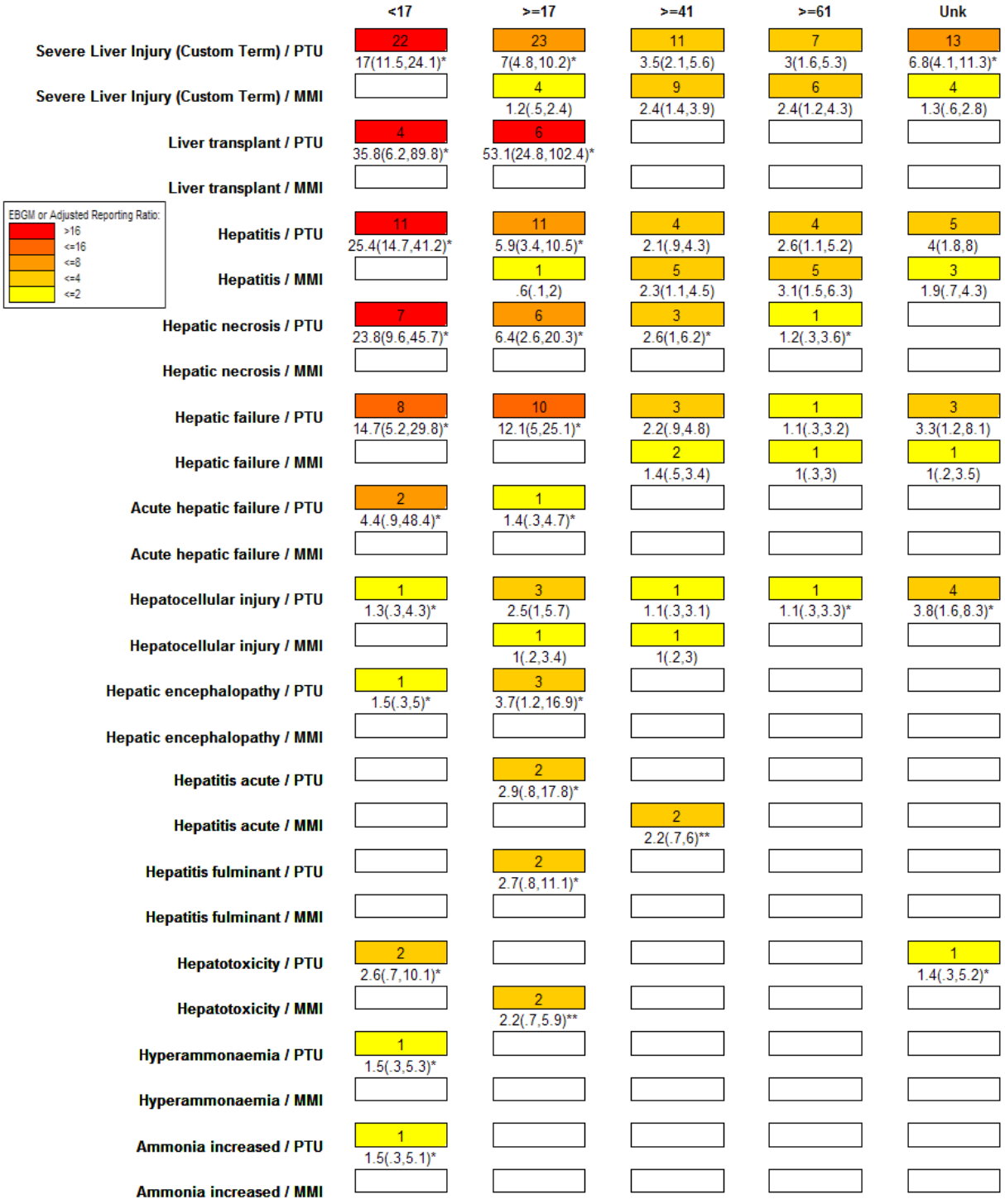


Rank	SOC	Term (PT)	EBGM
1	Hepat	Hepatitis	25.375
2	Hepat	Hepatic necrosis	23.838
3	Hepat	Hepatic failure	14.747
4	Hepat	Jaundice cholestatic	7.759
5	Hepat	Hyperbilirubinaemia	6.695
6	Hepat	Acute hepatic failure	4.408
7	Hepat	Autoimmune hepatitis	3.646
8	Hepat	Jaundice	3.397
9	Hepat	Cholestasis	2.909
10	Hepat	Hepatotoxicity	2.585
11	Hepat	Hepatic function abnormal	2.080
12	Hepat	Hepatomegaly	1.939
13	Hepat	Hepatitis cholestatic	1.566
14	Hepat	Hepatocellular injury	1.293

Rank	SOC	Term (PT)	EBGM
1	Hepat	Hepatic function abnormal	1.006

Figure 4 shows the MGPS values for these two drugs across age groups. The EBGM values for severe liver injury with propylthiouracil were higher in younger (≤17 yr) than older individuals (>17 yr), regardless of whether the individual severe preferred terms or the severe liver injury custom term grouping these individual terms was used. The side-by-side comparison of overlapping/non-overlapping status of (EB05, EB95) intervals for severe liver injury adverse events for all age groups showed them to be higher with propylthiouracil than methimazole, and the propylthiouracil EBGM values were higher and did not overlap with the methimazole values.

Figure 4: Paired Empirical Bayesian Geometric Mean (EBGM) values by drug and age group for the custom term that defines severe liver injury and for the individual preferred terms included in the custom term. Note the signal strength (EBGM value by color and by the annotation below each minigraph also containing the EB05-EB95 values) and the count of reports (number within each minigraph). * = Events with higher and non-overlapping EBGM values for propylthiouracil (PTU) than for methimazole (MMI) or events reported only with propylthiouracil; ** = events with higher and nonoverlapping EBGM values for methimazole than propylthiouracil or events reported only with methimazole. A higher EBGM score indicates a higher strength of association of the adverse event with the drug[22].**Error! Bookmark not defined.**


EBGM or Adjusted Reporting Ratio:
>16
<=16
<=8
<=4
<=2
<17
>=17
>=41
>=61
Unk
Severe Liver Injury (Custom Term) / PTU
22
17(11.5,24.1)*
23
7(4.8,10.2)*
11
3.5(2.1,5.6)
7
3(1.6,5.3)
13
6.8(4.1,11.3)*
Severe Liver Injury (Custom Term) / MMI
4
1.2(.5,2.4)
9
2.4(1.4,3.9)
6
2.4(1.2,4.3)
4
1.3(.6,2.8)
Liver transplant / PTU
4
35.8(6.2,89.8)*
6
53.1(24.8,102.4)*
Liver transplant / MMI
Hepatitis / PTU
11
25.4(14.7,41.2)*
11
5.9(3.4,10.5)*
4
2.1(.9,4.3)
4
2.6(1.1,5.2)
5
4(1.8,8)
Hepatitis / MMI
1
.6(.1,2)
5
2.3(1.1,4.5)
5
3.1(1.5,6.3)
3
1.9(.7,4.3)
Hepatic necrosis / PTU
7
23.8(9.6,45.7)*
6
6.4(2.6,20.3)*
3
2.6(1,6.2)*
1
1.2(.3,3.6)*
Hepatic necrosis / MMI
Hepatic failure / PTU
8
14.7(5.2,29.8)*
10
12.1(5,25.1)*
3
2.2(.9,4.8)
1
1.1(.3,3.2)
3
3.3(1.2,8.1)
Hepatic failure / MMI
2
1.4(.5,3.4)
1
1(.3,3)
1
1(.2,3.5)
Acute hepatic failure / PTU
2
4.4(.9,48.4)*
1
1.4(.3,4.7)*
Acute hepatic failure / MMI
Hepatocellular injury / PTU
1
1.3(.3,4.3)*
3
2.5(1,5.7)
1
1.1(.3,3.1)
1
1.1(.3,3.3)*
4
3.8(1.6,8.3)*
Hepatocellular injury / MMI
1
1(.2,3.4)
1
1(.2,3)
Hepatic encephalopathy / PTU
1
1.5(.3,5)*
3
3.7(1.2,16.9)*
Hepatic encephalopathy / MMI
Hepatitis acute / PTU
2
2.9(.8,17.8)*
Hepatitis acute / MMI
2
2.2(.7,6)**
Hepatitis fulminant / PTU
2
2.7(.8,11.1)*
Hepatitis fulminant / MMI
Hepatotoxicity / PTU
2
2.6(.7,10.1)*
1
1.4(.3,5.2)*
Hepatotoxicity / MMI
2
2.2(.7,5.9)**
Hyperammonaemia / PTU
1
1.5(.3,5.3)*
Hyperammonaemia / MMI
Ammonia increased / PTU
1
1.5(.3,5.1)*
Ammonia increased / MMI

The strong hepatoxicity signals with propylthiouracil were expected. The surprising finding was that methimazole, a drug used interchangeably in the same indication, was safer. Propylthiouracil has been restricted in practice to pregnant women who need to be treated in the first trimester. [22]

Safety report databases at FDA

Due to the unique analytic needs stemming from both product type characteristics and product type-specific regulatory authorities, no single adverse event database exists at the FDA for all products. Table 2 summarizes those FDA safety report databases for which data mining is used.

Table 2. Data mining of safety reports (reports of adverse events, injury, death, use errors, and hazardous product quality) received by FDA, by type of product, database characteristics, and data mining method. Databases that are too small for data mining were excluded.

	Database features as of Spring 2014			**Data mining method**	
Product type	**Current # reports received**	**Database start date**	**Cumulative # of reports**	**Stage of use**	**Method or tool**
Drugs	770,000 in 2013	1968	>7,000,000	Routine	MGPS with Empirica Signal™
				Develop-mental	V*ae*TM
				Develop-mental	MGPS with MASE™
				Develop-mental	GIS
Medical devices	670,000 in 2013	1991	3,300,000	Develop-mental	CPA
				Develop-mental	GIS
Vaccines	35,000 to 40,000	1990	200,000	Routine	• V*ae*TM • Network Analysis
				Develop-mental	AENA
				Develop-mental	GIS
Foods, Cosmetics, and Dietary Supplements	6,000 in 2013	2002	40,500	Routine	MGPS with Empirica Signal™
				Develop-mental	GIS

Product type	Database features as of Spring 2014			Data mining method	
	Current # reports received	**Database start date**	**Cumulative # of reports**	**Stage of use**	**Method or tool**
Animal drugs and devices	75,000	1991	400,000		PRR and MGPS with PV Analyzer™
				Develop-mental	GIS

*Table 2 Notes:

- "Drugs" includes the following products intended for human use: prescription drugs, over the counter drugs, homeopathic drugs, human cellular products, blood derivatives, and products that are a combination of medical device and drug.
- "Medical Devices" include products, that are a combination of medical device and biologic, that are not in the "Drugs" category.

The place of data mining in assessment of safety reports

Data mining analyses are used to detect potential signals and generate related hypotheses, but cannot be used in isolation to establish causality. Many possible reasons have been found for the statistical association between a product and an event other than a direct causal relationship [24] (e.g., recent questions about Pradaxa® [25]). Hands-on case reviews, analysis of other data sources (e.g. FDA regulatory databases, World Health Organization drug safety report database, [26] public scientific literature, and public knowledge databases [27, 28, 29]) and further epidemiologic assessments [25], [30] are necessary to characterize the clinical and public health significance of "signals" generated by data mining analyses. [2]

When the evidence of a new safety issue is compelling, FDA may take regulatory action (such as a product recall or changes in product labeling) and is responsible for informing the public of these actions, along with any firm-initiated communications.

PAST SUCCESSFUL MINING OF SAFETY REPORT DATABASES

Mining FDA safety report databases has identified important safety issues in recent years.

The first vaccine safety signal detected with use of MGPS alone was febrile seizures associated with Fluzone® 2010-11 administration in young children. [31] The signaling threshold, database

restrictions, adjustment, and baseline data mining were strategies adopted *a priori* to enhance the specificity of the data mining analyses of the 2010–11 influenza vaccine.

Data mining has assisted in evaluation of many important drug safety signals, including the associations of pituitary tumors with atypical antipsychotics, [32] pathological gambling with parkinsonian therapy, [23] and pancreatitis with atypical antipsychotics and valproic acid. [7] Even data for older drugs may contain hidden signals of toxicity elicited by data mining, as was the case for hepatotoxicity associated with propylthiouriacil. [22] The importance of evaluating other data in conjunction with signals identified by data mining was exemplified in the evaluation of amyotrophic lateral sclerosis associated with statins. [33]

Mining dietary supplement safety report data identified unusual levels of liver toxicity associated with the weight-loss dietary supplement Hydroxycut®. Further investigation of the clinical records of the patients with liver damage who took Hydroxycut® confirmed that the relative timing of Hydroxycut® use and liver damage was consistent with causality, and in most cases, no other cause of liver damage could be found. [34] Hydroxycut® was voluntarily recalled from the market in May 2009 due to hepatic toxicity. [35] Hydroxycut® was subsequently reformulated and remarketed.

Retrospective data mining of the MAUDE database showed that safety signals associated with an implantable cardioverter defibrillator (ICD), could have been detected in March 2006. [36] Using traditional methods, the signal of lead fracture and inappropriate shock events related to Sprint Fidelis® leads was actually detected 10 months later, in January 2007. The manufacturer announced a voluntary market withdrawal in October 2007.

These examples highlight the important role data mining has played in product safety report surveillance at the FDA.

DATA MINING METHODS APPLIED TO OTHER TYPES OF DATA

Encouraged by the success of using data mining methods for safety report analysis, FDA experts have started to apply the techniques to other types of data, summarized in Table 3.

Table 3. Types of data, and the data mining methods used for them at FDA.

Type of data	Stage of use of data mining	Data mining method or tool	Data mining purpose
MEDLINE®	Develop-mental	Disproportionality analysis	Find drug-adverse event signal pairs

Type of data	Stage of use of data mining	Data mining method or tool	Data mining purpose
Medical literature	Develop-mental	Linguamatics I2E natural language processing; using chemical structure information from the medical literature	Study clinical safety
		G-VISR (Georgetown Vaccine Information and Safety Resource) tool	Collect molecular and adverse event information
Medical device documents	Develop-mental	SARF semantic text mining	• Search within any number of repositories • Screen for massive lists of items within repositories
Clinical study data in drug applications	Routine	Empirica Study™ creation of a wide set of automatically generated analytical outputs and tailor-made, reusable tables and graphs	Save reviewers from having to create the tables and graphs
Social media	Develop-mental	MedWatcher Social; uses standard product and adverse event dictionaries	Detect adverse events related to medical products
Tobacco documents	Develop-mental	Topic modeling methods	Characterize documents and estimate topics covered by the documents
Questions received at the CFSAN Call Center	Develop-mental	SAS™ data step programming and SAS™ text mining node	Categorize and group the predominant types of questions
	Routine	SAS Enterprise Miner™	Maintain standardized data fields

Disproportionality analysis

CDER has partnered with the National Library of Medicine (NLM) to identify disproportionate reporting of drug-adverse event pairs in MEDLINE®, the National Library of Medicine (NLM)'s publicly available database of over 20 million biomedical abstracted articles and citations. Experts in cognitive science and linguistics from NLM have mapped the medical subject headings (MeSH) terms [37] used for indexing of citations in MEDLINE® with adverse events terminology in the Medical Dictionary for Regulatory Activities (MedDRA). [38] MeSH terms related to drug names have been mapped to the Anatomical Therapeutic Chemical (ATC) Classification System and RxNorm. [28]

CDER has applied Empirica Study™ and other software packages such as SAS JMP™ and JReview™ to analyze drug clinical trial data in either New Drug Applications or supplemental applications. Empirica Study™ interfaces with data that conforms to the standardized Study Data Tabulation Model (SDTM) of the Clinical Data Interchange Standards Consortium (CDISC) data standards to create a wide set of automatically generated analytical outputs and tailor-made, reusable tables and graphs. These outputs have helped reviewers to more efficiently analyze potential safety issues in clinical trial data of drugs approved by the FDA. [39]

Text mining

CDER has also explored text mining using Linguamatics™ I2E software to study clinical safety based on chemical structure information contained in medical literature. Linguamatics™ I2E enables custom searches using natural language processing (NLP) to interpret unstructured text. The ability to predict clinical safety based on chemical structures is becoming increasingly important, especially when adequate safety data are absent or equivocal. [40]

A semantic text mining (STM) tool is being researched with a view to creating a scalable, secure, industrial-scale, and flexible framework for the widest possible variety of text mining applications to reside upon. The Search and Retrieval Framework (SARF), which was developed by CDRH, is now able to both search within any number of available repositories and screen for massive lists of items within those repositories. SARF includes state-of-the-art ontologies maintained by the National Library of Medicine (NLM) and by the FDA along with general-purpose dictionaries. Any number of new dictionaries can be added and selected by the user.

For vaccines, CBER is working with the Innovation Center for Biomedical Informatics at the Georgetown University on the development of G-VISR (Georgetown Vaccine Information and Safety Resource) tool. G-VISR mines the biomedical literature and existing databases to collect molecular and adverse event information related to particular vaccines.

The Office of the Chief Scientist is studying the utility of signal detection from social media. MedWatcher Social is an exploratory data mining tool to detect adverse events related to medical products, using publicly available data on social media (Twitter, Facebook, health-related web blogs) to curate and map health information to standard product and adverse event dictionaries.[41]

MedWatcher Social has the potential to incorporate logarithmic Internet search terms in the near future.

The SAS Enterprise Miner™ (SASEM™) specialized text mining software package was recently used to perform text mining of consumer, industry, and governmental questions received by the Call Center in the Center for Food Safety and Nutrition (CFSAN). The combination of SAS™ data step programming and the SAS™ text mining node was useful in categorizing and grouping the predominant types of inquiries received. ext mining plays an important role in maintaining standardized data fields at the CFSAN Call Center.

CDRH's Office of Science and Engineering Laboratories (OSEL) is also developing, for medical device documents, behavioral linguistic methods to analyze free text fields to extract manufacturer reporting patterns, and vector, matrix, and free-space approaches to text association.

Topic modeling

CTP is using topic modeling to characterize document content based on key terms and to estimate topics contained within documents. It can also be used to estimate and identify topics from the document, word and phrase content, and cluster documents. For example, documents associated with the topic "menthol" would compose one cluster. Documents on menthol describing patterns of usage in "youth" would then be a subset of this more general cluster.

Another useful application of topic modeling at CTP is the evaluation of evidence associated with a Modified Risk Tobacco Product Application (MRTPA). Although many of the documents associated with an MRTPA are of a historical nature, the context of this assessment is the pre-market setting; that is, before FDA issued an order to market.

CTP is collaborating with NCTR to develop software to implement Latent Dirichlet Allocation (LDA). LDA is a method used in text mining to automatically identify topics that are contained in disparate text.

Other specific topic modeling techniques being explored for tobacco documents include:

- k-methods (k-means and k – nearest neighbor)
- hierarchical clustering
- latent variable Latent Semantic Analysis (LSA)
- Probabilistic Latent Semantic Analysis (PLSA).

ADVANTAGES AND CHALLENGES OF DATA MINING

Advantages of mining safety report databases

FDA has experienced the following advantages of data mining:

- *Standard Processes.* Historically, manual analyses (whether in generating a specific hypothesis, selecting the event codes to analyze, or selecting a case series or cohort by chart review) raised concerns regarding the accuracy, subjectivity, reproducibility, and interpretation of the data used for conducting the analyses. In contrast, because data mining is automated, the outputs produced are systematic and statistically "objective," given the limitations of the data.
- *Simultaneous Analysis.* Data mining calculations are made without *a priori* hypotheses for every product-event combination across an entire database at once.
- *Efficiency*. The signal scores for all the product-event pairs are computed in minutes, much faster than manually requesting traditional computerized exploratory analyses.
- *Prioritization of Investigating Signals*. Data mining enables much easier prioritization of investigating signals based on the seriousness of the event; the magnitude of data mining scores; the redundancy of clusters of patterns for the product, product class and/or indication; and the number of collateral (similar) adverse event terms.
- *Automated support of further signal investigation*.
    - "Drill down" capabilities enhance manual exploration.
    - Stratification and sub-setting.
    - Observation of signals over time.
    - Identification of complex interdependent factors (e.g., concomitant products and/or diseases).
    - Facilitation of the study of product interactions by automatically calculating unusual reporting patterns for patients using multiple products (e.g., a drug for hypertension and a pacemaker).
    - Transparency, replication, and collaboration are fostered by detailed audit trails.
    - Identification and correction of data errors
    - Facilitation of the planning of database and analytic improvements
    - Support for understanding of the biological plausibility of signals by incorporating reference datasets regarding chemistry and physiology.

Challenges and data-mining mitigations related to safety report databases

Challenges, inherent in safety report databases, that limit the interpretability of signals have already been discussed elsewhere [2, 7, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52] and include:

- *Missing, incorrect, or vague information.* For our current data mining tools to run efficiently, the data must be of high-quality. Errors and inconsistencies are the most common data quality issues and must be kept to a minimum. Common types of quality errors include coding, formatting, and data entry mistakes. When critical information contained in narratives must be manually extracted by reviewers or data entry personnel, important information can be inadvertently missed because of human error.
- *Separate reports about the same incident.* Information about the same event may be submitted by the patient's physician, pharmacist, nurse, attorney, and/or the patient. Multiple companies may also submit information on the same case if the patient was exposed to several concomitant products manufactured by different companies. There may also be a series of follow-up reports for the same case as additional information becomes available. It is important to consider the best representative reports (for cases having multiple sources and updated reports) before analyzing cases. In addition, a report may mention products that span the oversight of various FDA Centers (e.g., a report mentioning a drug, device, and a nutritional supplement) and could be entered into several different databases for analysis from different reviewers with differing perspectives.
- *Event may be due to the treated condition, another condition, or another product.* This is a common concern for patients with multiple, complicated medical conditions for which they are using multiple products. *Underreporting.* Underreporting could be due to lack of recognition of a possible product-event association, lack of awareness of reporting expectations or process, fear of litigation, or reporting to another public health organization instead of FDA. [42, 47, 48, 49, 50, 51] This underreporting may vary according to the type of product, the seriousness of the event, the population using the product, the product's time on the market, and other factors. It has been estimated that 94% of adverse drug reactions go undetected by spontaneous reporting systems. [42] Furthermore, common events, such as heart attacks, may be under-recognized as adverse events, such as occurred with Vioxx (rofecoxib) [46] before epidemiology studies were done.
- *Over-reporting*. Over-reporting can be due to media publicity,[43, 44] litigation,[45] or the product being newly marketed. [52]
- *Timeliness of reporting and processing*

FDA's disproportionality and CPA tools work best on databases that use standard terms for the product, event, and co-variables, such as age. Although much of the standardization is done manually, text mining is being investigated as a tool to assist with standardization and use text fields to enhance the content of coded fields. The joint use of clustering apparently related products, apparently related events, and standard references for products and events, has helped analysts address the problems of incorrect or vague information in reports.

Challenges related to the application of data mining to safety reports

Specific data mining methodologies and interpretation of signals requires database-specific understanding of:

- *Acceptance of foreign (versus domestic) reports, with different reporting requirements*. Foreign reports in the FDA databases are likely to be serious, unlabeled events, whereas U.S. reports usually contain both labeled and unlabeled events, regardless of seriousness. Labeled adverse events are events seen in the premarket clinical review of drugs, devices, and vaccines and listed by the manufacturer on the label as a potential risk of product use. Databases with a high proportion of foreign to domestic reports may elicit more "signals" if the foreign reports contain a high proportion of serious, unlabeled events. In addition, products may have different indications for use in different countries.
- *Changing reporting requirements over time*
- *Changing coding dictionaries for products and events*. The results can include discrepant product names and/or events.
- *Changing data entry and coding processes*
- *Inconsistent database structure architecture*
- *Malicious reporting and spam*

It is critical for the FDA's data mining disproportionality tools that the data be as "clean" and as standardized as possible. Converting data to a standard form and/or recoding the data can be expensive and error prone. For field variables such as sex, date, product, and event, it is essential that the values used to designate these variables be precisely defined and consistently used. Furthermore, the text fields in safety reports can contain very diverse content, such as laboratory tests, blood pressure and weight measurements, and past medical history. Although data attributes are often described by a single word or text string, a rigorous terminology standard requires:

- a concept definition
- a code for the concept
- a list of labels (terms) associated with the concept

A concept definition is a detailed description of the concept associated with a specific concept code. For example, a variety of labels is currently associated with PubMed code CID 2244 (aspirin) that differ in spelling, letter case, and spacing of product names. An informed human being can readily tell that "Aspirin," "aspirin," "acetyl salicylic acid", "acetylsalicylic acid", and "ASA" all refer to the same active ingredient. However, a computer cannot make this distinction (unless specifically programmed to do so) and will therefore treat these labels as different active ingredients even though the active ingredient is the same.

Signal thresholds are adjusted to account for the severity of the adverse event related to the product and the severity of the condition for which the product is being used. For example, the threshold for evaluating a safety issue for a drug used to treat cancer would be different than the threshold for a drug used to treat acne. [2, 7, 53]

Additional challenges specific to interpreting signals generated by safety reports data mining include:

- All of the reports represent a reporter's concern that there is a product-event relationship; signals do not reflect actual rates of events per product use.
- The signals are database-specific. The contents of each database are functions of separate regulatory authorities, rather than simple inherent affinities.

Advantages and challenges of using data mining for other data types

Data mining of other sources, such as medical literature, electronic health records and social media, shares many of the challenges related to safety reports data. The quality of data in these sources can be better or worse, depending on the structure of the database and the training of those who enter the data, varying from presumably high quality Medical Subject Headings (MeSH®) indexes of MEDLINE® [37] to social media (e.g. Facebook) [41] and web blogs.

THE FUTURE OF DATA MINING AT FDA

Analytic challenges will continue to grow with the addition of new surveillance data sources and the development of new methods of submitting spontaneous reports, such as web-based and mobile applications. It will be important for the FDA to structure its IT systems so that data can be submitted, retrieved, processed, and evaluated in a standardized manner.

There will be vast increases and changes in surveillance data that will be reported to and available to FDA in the near future. These include electronic health records [54] and claims, [55] personal health records, [56] standards for health data, [54] data from Federal and private sector mobile devices for tracking health, [57, 58, 59] and data from social websites (blogs, patient advocacy group sites, and search term logs). [41, 60, 61]

Outside work has shown that these data sources can be of value in post-market safety surveillance and other related fields; [41, 51, 55, 60] FDA would like to validate their utility for surveillance of FDA-regulated products.

Further development and implementation of an advanced and integrated safety data mining system supported by appropriately experienced personnel will be essential for better informed decision making and risk management of product safety issues in real time. Specific desired data mining capabilities include:

- Scalability to accommodate growing databases.
- Further advanced natural language processing and text mining to automatically and accurately extract meaning from narratives in all sorts of databases.
- Either data processing that is very quick, or methods that require less data processing, to move surveillance closer to real time.
- Additional advanced visual analytics with more advanced drill down functions coupled to context information across multiple data resources.

- Complete reference databases for topics including:
    - Product characteristics
    - Event characteristics
    - Physiology
    - Toxicology

- More transparent human-readable audit trails to enable analysts to more efficiently communicate and validate each other's selection criteria, results, and interpretation.

As a result, researchers and policy makers will be better equipped to understand the limitations and biases of the data, leading to more informed decisions regarding FDA-regulated product safety.

Data mining holds promise for other FDA work, including:

- FDA field work. Potential uses include exploring trends in safety, inspection, and recalls data so that field managers can more effectively align available personnel and resources to have the greatest impact on public health. Data mining could also assist in enforcement coordination among district and headquarters personnel.
- Pre-approval safety reviews and efficacy evaluations of products.
- Information contained in tobacco health documents, including legal documents and research reports on a range of topics, including:
    - dose response relationships
    - chemosensory effects
    - neurobiology of dependence
    - menthol-nicotine interactions
    - product-related interactions
    - advertising-related perceptions
    - marketing strategies
    - switching rates
    - initiation and cessation rates

The FDA Data Mining Council, composed of the authors and other interested staff from across FDA, promotes the improvement of data mining to support FDA's mission of protecting and promoting public health. We advocate sharing expertise among other government agencies, academia and private sector companies to increase knowledge about data mining and improve data analysis.

ABOUT THE FDA DATA MINING COUNCIL

In response to the need to develop a FDA wide data mining collaboration and strategy, the FDA Data Mining Council (DMC) was formed in 2007. The Council serves as a forum for FDA scientists to share their experiences and challenges in analyzing data contained in the vast databases the FDA maintains, as well as to discuss new methods for such analyses. The DMC has explored the following:

- how Centers collect and analyze data
- novel sources of data
- data standardization
- data mining tools and statistical principles
- text mining of unstructured data
- data visualization techniques

DMC members discuss their research on a regular basis and provide information and assistance within the FDA regarding the broad areas listed above. The DMC is collaborative and explores methods and best practices recommended by experts from other federal agencies, industry, and academia—all of whom have analogous experience in knowledge discovery through various data mining approaches.

The contact for this page isHesha Duggirala, PhD, DMC Chair, at Hesha.Duggirala@fda.hhs.gov.

References

[1] Reducing and Preventing Adverse Drug Events To Decrease Hospital Costs: Research in Action, Issue 1. March 2001. Agency for Healthcare Research and Quality, Rockville, MD. http://www.ahrq.gov/legacy/qual/aderia/aderia.htm. Accessed December 2014.

[2] Guidance for Industry. Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment. Food and Drug Administration, US Department of Health and Human Services. March 2005. http://www.fda.gov/downloads/regulatoryinformation/guidances/ucm126834.pdf. Accessed Dec 2014.

[3] Waller PC, Evans SJ. A model for the future conduct of pharmacovigilance. *Pharnaccoepidemiol Drug Saf*. 2003 Jan-Feb;12(1):17-29.

[4] Finney DJ. Systemic signalling of adverse reactions to drugs. *Methods Inf Med*. 1974 Jan;13(1):1-10.

[5] Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* 2001 Oct-Nov;10(6):483-6.

[6] Bate A, Evans, S. Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiol and Drug Saf* 2009 Jun;18(6):427-36. doi: 10.1002/pds.1742.

[7] Szarfman A, Tonning JM, Doraiswamy PM. Pharmacovigilance in the 21st century: new systematic tools for an old problem. *Pharmacotherapy* 2004 Sep;24(9):1099-104.

[8] Dumouchel W. Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System, *American Statistician* 1999; 53(3):177-190.

[9] Chirtel S. Disproportionality analysis for detection of food adverse events. Presented at Eastern North American Region of the International Biometric Society, Washington DC, Spring 2012, in Statistical Applications in Food safety section.

[10] Kass-Hout TA, Xu Z, McMurray P, et al. Application of change point analysis to daily influenza-like illness emergency department visits. *J Am Med Inform Assoc* 2012 Nov-Dec;19(6):1075-81. doi: 10.1136/amiajnl-2011-000793.

[11] Kass-Hout TA, Xu, Z. Change point analysis. https://sites.google.com/site/changepointanalysis. Accessed Feb 2015.

[12] Edwards AW, Cavalli-Sforza LL. A method for cluster analysis. *Biometrics* 1965 Jun;21:362-75.

[13] Auger IE, Lawrence CE. Algorithms for the optimal identification of segment neighborhoods. *Bull Math Biol* 1989;51(1):39-54.

[14] Bai J, Perron,P. Estimating and testing linear models with multiple structural changes. *Econometrica* 1998 Jan;66(1):47-78.

[15] Killick R, Fearnhead P, Eckley IA. Optimal detection of changepoints with a linear computational cost. *JASA* 2012;107(500):1590-1598.

[16] Killick R, Eckley IA. Changepoint: An R package for changepoint analysis. 2012. http://www.lancs.ac.uk/~killick/Pub/KillickEckley2011.pdf. Accessed 2/14/2015.

[17] Botsis T, Buttolph T, Nguyen M, et al. Vaccine adverse event Text Mining (VaeTM) system for extracting features from vaccine safety reports. *J Am Med Inform Assoc* 2012 Nov-Dec;19(6):1011-8. doi: 10.1136/amiajnl-2012-000881.

[18] Moore PW, Burkhart KK, Jackson D. Drugs highly associated with infusion reactions reported using two different data-mining methodologies. *J Blood Disorders Transf* 2014;5:195. doi: 10.4172/2155-9864.1000195.

[19] FDA's Geographic Information System. http://www.fda.gov/AboutFDA/CentersOffices/OC/OfficeoftheCounselortotheCommissioner/ucm227114.htm. Last updated 4/2/2012. Accessed Feb 2015.

[20] Ball R and Botsis T. Can network analysis improve pattern recognition among adverse events following immunization reported to VAERS? *Clin Pharmacol Ther* 2011 Aug;90(2):271-8. doi: 10.1038/clpt.**2011**.119.

[21] Botsis T, Scott J, Goud R, et al. Novel algorithms for improved pattern recognition using the US FDA adverse event network analyzer. *Stud Health Technol Inform*. 2014; 205:1178-82.

[22] Rivkees SA, Szarfman A. Dissimilar hepatotoxicity profiles of propylthiouracil and methimazole in children. *J Clin Endocrinol Metab* 2010 Jul;95(7):3260-7. doi: 10.1210/jc.2009-2546.
[23] Szarfman A, Doraiswamy PM, Tonning JM, et al. Association between pathologic gambling and parkinsonian therapy as detected in the Food and Drug Administration Adverse Event database. *Arch Neurol*. 2006 Feb;63(2):299-300.
[24] Almenoff J, Tonning JM, Gould AL, et al. Perspectives on the use of data mining in pharmacovigilance. *Drug Saf* 2005;28(11):981-1007.
[25] Pradaxa (dabigatran): Drug Safety Communication - Lower Risk for Stroke and Death, but Higher Risk for GI Bleeding Compared to Warfarin. Posted 05/13/2014. http://www.fda.gov/safety/medwatch/safetyinformation/safetyalertsforhumanmedicalproducts/ucm397179.htm. Accessed Jan 2015.
[26] Vigibase. World Health Organization. Last updated 12/19/2014. http://who-umc.org/DynPage.aspx?id=98082&mn1=7347&mn2=7252&mn3=7322&mn4=7326. Accessed Feb 2015.
[27] TOXNET databases. U.S. National Library of Medicine. http://toxnet.nlm.nih.gov/. Accessed Feb 2015.
[28] Medical terminologies at NLM. U.S. National Library of Medicine. Last reviewed 12/2/2013. http://www.nlm.nih.gov/medical-terms.html. Accessed Feb 2015.
[29] DAILYMED. U.S. National Library of Medicine. http://dailymed.nlm.nih.gov/dailymed/index.cfm. Accessed Feb 2015.
[30] Welcome to Mini-Sentinel. Food and Drug Administration. Last updated 10/15/2014. http://www.mini-sentinel.org/. Accessed Feb 2015.
[31] Martin D, Menschik M, Bryant-Genevier M, et al. Data mining for prospective early detection of safety signals in the Vaccine Adverse Event Reporting System (VAERS): a case study of febrile seizures after a 2010-2011 seasonal influenza virus vaccine. *Drug Saf*. 2013 Jul;36(7):547-56. doi: 10.1007/s40264-013-0051-9.
[32] Szarfman A, Tonning JM, Levine JG, et al. Atypical antipsychotics and pituitary tumors: a pharmacovigilance study. *Pharmacotherapy*. 2006 Jun;26(6):748-58.
[33] Colman E, Szarfman A, Wyeth J, et al. An evaluation of a data mining signal for amyotrophic lateral sclerosis and statins detected in FDA's spontaneous adverse event reporting system. *Pharmacoepidemiol Drug Saf* 2008 Nov;17(11):1068-76. doi: 10.1002/pds.1643.
[34] Fong T-L, Klontz KC, Canas-Coto A, et al. Hepatoxicity due to Hydroxycut®: a case series. *Am J Gastroenterol* 2010 July; 105(7): .doi:10.1038/ajg.2010.5.
[35] Warning on Hydroxycut. FDA. 01/20/2015 http://www.fda.gov/ForConsumers/ConsumerUpdates/ucm152152.htm. Accessed Feb 2015.
[36] Duggirala HJ, Herz ND, Caños DA, et al. Disproportionality analysis for signal detection of implantable cardioverter-defibrillator-related adverse events in the Food and Drug Administration Medical Device Reporting System. *Pharmacoepidemiol Drug Saf* 2012 Jan;21(1):87-93. doi: 10.1002/pds.2261.
[37] Fact Sheet: Medical Subject Headings (MeSH®). U.S. National Library of Medicine. September 12, 2013. http://www.nlm.nih.gov/pubs/factsheets/mesh.html. Accessed Feb 2015.
[38] Welcome to MedDRA. ICH Steering Committee. http://www.meddra.org/. Accessed Feb 2015.
[39] Szarfman A. Medical officer's consultative reanalysis of the febrile neutropenia studies of NDA 50-679. http://www.fda.gov/downloads/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/DrugSafetyInformationforHeathcareProfessionals/UCM201520.pdf . Accessed Dec 2014.
[40] Botsis T, Ball R. Automating case definitions using literature-based reasoning. *Appl Clin Inform*. 2013 Oct 30;4(4):515-27. doi: 10.4338/ACI-2013-04-RA-0028.
[41] Freifeld CC, Brownstein JS, Menone CM, et al. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. *Drug Saf*. 2014 May;37(5):343-50. doi: 10.1007/s40264-014-0155-x.
[42] Hazell L, Shakir SA. Under-reporting of adverse drug reactions: a systematic review. *Drug Saf*. 2006;29(5):385-96.
[43] Waller PC. Measuring the frequency of adverse drug reactions. *Br J Clin Pharmac* 1992 Mar;33(3):249-252.
[44] Meinzinger MM, Barry WS. Prospective study of the influence of the media on reporting medical events. *Therapeutic Innovation & Regulatory Science*. 1990 Jul;24(3):575-577.

[45] McAdams M, Staffa J, Dal Pan G. Estimating the extent of reporting to FDA: a case study of statin-associated rhabdomyolysis. *Pharmacoepidemiol Drug Saf*. 2008 Mar,17(3):229-39. doi: 10.1002/pds.1535.
[46] Graham DJ, Campen D, Hui R, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet*. 2005 Feb 5-11;365(9458):475-81.
[47] Bright RA. Surveillance of adverse medical device events. IN: Brown SL, Bright RA, Tavris DR, eds. *Medical Device Epidemiology and Surveillance*. John Wiley & Sons, Ltd., 2007:43-61.
[48] Balka E, Doyle-Waters M, Lecznarowicz D, et al. Technology, governance, and patient safety: Systems issues in technology and patient safety. *Int J Med Inform*. 2007 Jun:76 Suppl 1:S35-47.
[49] Samore MH, Evans RS, Lassen A, et al. Surveillance of medical device-related hazards and adverse events in hospitalized patients. *JAMA* 2004 Jan 21; 291(3):325-334.
[50] Medical devices: early warning of problems is hampered by severe underreporting. US General Accounting Office. GAO/PEMD 87-1; 1987.
[51] Hefflin B, Gross T, Schroeder T. Estimates of medical device-associated adverse events from emergency departments. *Am J Prev Med* 2004; 27(3):246-253.
[52] Tsong Y. Comparing reporting rates of adverse events between drugs with adjustment for year of marketing and secular trends in total reporting. J Biopharm Stat 1995; 5:95-114.
[53] Bright RA, Nelson RC. Automated support for pharmacovigilance: a proposed system. *Pharmacoepidemiol Drug Saf.* 2002 Mar;11(2):121-125.
[54] Update on the adoption of health information technology and related efforts to facilitate the electronic use and exchange of health information. Report to Congress. Office of the National Coordinator for Health Information Technology, US Department of Health and Human Services. Oct 2014. http://www.healthit.gov/sites/default/files/rtc_adoption_and_exchange9302014.pdf . Accessed Feb 2015.
[55] Zhan C, Kaczmarek R, Loyo-Berrios N, et al. Incidence and short-term outcomes of primary and revision hip replacement in the United States. *J Bone Joint Surg Am*. 2007 Mar; 89(3):526-33.
[56] Deering MJ. Issue brief: patient-generated health data and health IT. Office of the National Coordinator for Health Information Technology, US Department of Health and Human Services. 12/20/2013. http://www.healthit.gov/sites/default/files/pghd_brief_final122013.pdf. Accessed Feb 2015.
[57] Sands DZ, Wald JS. Transforming health care delivery through consumer engagement, health data transparency, and patient-generated health information. *Yearb Med Inform*. 2014; 9(1):170-176. doi: 10.15265/IY-2014-0017.
[58] HealthData.gov. US Department of Health and Human Services. http://www.healthdata.gov/. Accessed Feb 2015.
[59] National Electronic Injury Surveillance System (NEISS). US Consumer Product Safety Commission. http://www.cpsc.gov/en/Research--Statistics/NEISS-Injury-Data/. Accessed Feb 2015.
[60] Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. *Nature*. 2009 Feb 19;457(7232):1012-4. doi: 10.1038/nature07634.
[61] White RW, Harpaz R, Shah NH, et al. Toward enhanced pharmacovigilance using patient-generated data on the internet. *Clin Pharmacol Ther*. 2014 Aug;96(2):239-246. Doi:10.1038/clpt.2014.77.